Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4

Year: 2023 | Citations: 302 | Authors: Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou

---

Abstract

Harnessing logical reasoning ability is a comprehensive natural language understanding endeavor. With the release of Generative Pretrained Transformer 4 (GPT-4), highlighted as"advanced"at reasoning tasks, we are eager to learn the GPT-4 performance on various logical reasoning tasks. This report analyses multiple logical reasoning datasets, with popular benchmarks like LogiQA and ReClor, and newly-released datasets like AR-LSAT. We test the multi-choice reading comprehension and natural language inference tasks with benchmarks requiring logical reasoning. We further construct a logical reasoning out-of-distribution dataset to investigate the robustness of ChatGPT and GPT-4. We also make a performance comparison between ChatGPT and GPT-4. Experiment results show that ChatGPT performs significantly better than the RoBERTa fine-tuning method on most logical reasoning benchmarks. With early access to the GPT-4 API we are able to conduct intense experiments on the GPT-4 model. The results show GPT-4 yields even higher performance on most logical reasoning datasets. Among benchmarks, ChatGPT and GPT-4 do relatively well on well-known datasets like LogiQA and ReClor. However, the performance drops significantly when handling newly released and out-of-distribution datasets. Logical reasoning remains challenging for ChatGPT and GPT-4, especially on out-of-distribution and natural language inference datasets. We release the prompt-style logical reasoning datasets as a benchmark suite and name it LogiEval.